Katsenou, A., Ntasios, T., Afonso, M., Agrafiotis, D., & Bull, D. (2018). Understanding video texture — A basis for video compression. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP 2017): Proceedings of a meeting held 16-18 October 2017, Luton, United Kingdom* (pp. 232-238). [8122252] Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/MMSP.2017.8122252

UNDERSTANDING VIDEO TEXTURE - A BASIS FOR VIDEO COMPRESSION

Angeliki V. Katsenou, Thomas Ntasios, Mariana Afonso, Dimitris Agrafiotis and David R. Bull

Department of Electrical and Electronic Engineering, University of Bristol, Bristol BS8 1UB, UK
{Angeliki.Katsenou, T.Ntasios, Mariana.Afonso, D.Agrafiotis, Dave.Bull}@bristol.ac.uk

ABSTRACT

Encoding spatio-temporally varying textures is challenging for standardised video encoders, with significantly more bits required for textured blocks compared to non-textured blocks. It is therefore beneficial to understand video textures in terms of both their spatio-temporal characteristics and their encoding statistics in order to optimize coding modes and performance. To this end, we examine the classification of video texture based on encoder performance. For this purpose, we employ spatio-temporal features and follow a two-step feature selection process by employing unsupervised machine learning approaches across the selected feature space. Finally, supervised machine learning approaches are applied on the set of the selected features that support classification prior to encoding with up to 95.1% accuracy. The results of this study will form the basis of a new informed approach to codec configuration and mode selection in both current and future encoders.

Index Terms— Video Texture, Textural Features, Video Compression, HEVC, Machine Learning.

I. INTRODUCTION

In the context of video compression, texture has been typically categorized into two different classes: static and dynamic [1]. However, the classification of dynamic textures in these categories is very generic, with a large range of diverse content being included in the same class, e.g. water and foliage. A recent statistical analysis of video textures in the context of the *High Efficiency Video Coding* (HEVC) standard [2] has shown that the video encoder handles these different types of texture very differently in terms of coding modes and bit rate [3], [4]. For example, for the homogeneous low resolution (256×256) video sequences in HomTex [5], in HEVC HM16.2, using random access mode and quantization level 27, requires twice the number of bits per pixel (bpp) for dynamic discrete textures compared to dynamic continuous textures and five times higher that for static textures. Due to these different spatio-temporal characteristics, different encoding modes are usually used for dynamic continuous textures (e.g. flowing water) than for dynamic discrete textures (e.g. foliage) or static textures (e.g. a camera panning over a carpet). This motivates us to examine more closely the different types of texture from uncompressed videos using their low-level features as a basis for classification, as this will provide important information about the coding performance for different types of texture. The outcomes of this study are expected to influence encoder design by facilitating encoding decisions and by introducing of new encoding modes.

Traditionally, all types of dynamic texture are included in a single category [6], [7]. Two of the first approaches that recognised that dynamic textures exhibit different types of spatio-temporal behaviours were reported in [8] and [9]. In [9], research on the definition of different spatio-temporal patterns was used to build a semantic texture classifier using categories such as fire, water, crowds etc. However, this work focuses on the recognition of the texture mainly for scene description and scene understanding applications. This work belongs to a much wider body of classification and recognition methods for dynamic textures, e.g. [10], [11], that consider only semantic categories of dynamic texture. None of these relate to video texture compression.

Renaud et al. [8], provide an explanation of the categorization of texture in the annotated DynTex dataset with multiple attributes among which are dynamic discrete, dynamic continuous and static. The same category definitions were used in our initial attempt to understand video texture for compression purposes in [3], where the analysis of encoding statistics revealed very different encoding decisions (e.g. prediction modes, *Coding Tree Unit* (CTU) partitioning) and performances (e.g. bits used for residual encoding) for the different texture types. Moreover, the hypothesis of the three main texture classes was verified by applying unsupervised learning methods on the extracted HEVC encoding statistics.

In this paper, first we independently explore the types of video texture from the perspective of their spatio-temporal characteristics and from the perspective of encoding decisions and performance statistics. Based on these outcomes we argue that it is meaningful to categorise content into the three classes (partially also verified by clustering the coding statistics in [3]) by extracting textural features from uncompressed videos. We justify this by applying *Expecta-*

THE WORK PRESENTED WAS SUPPORTED BY THE "MARIE SKŁODOWSKA-CURIE ACTIONS - THE EU FRAMEWORK PROGRAMME FOR RESEARCH" PROJECT PROVISION.

tion Maximization (EM) clustering on the extracted spatio-temporal features. Next, based on the extracted features and by adopting a supervised learning algorithm, we build and train two video texture classifiers. The first of these relates features from uncompressed content to texture classes as defined by experts, while the second relates features from uncompressed content to classes directly representing patterns of HEVC HM encoding statistics. To the best of our knowledge, this is the first approach that uses features extracted from uncompressed content to predict encoder performance via texture classification in dynamic discrete, dynamic continuous, and static textures.

The remainder of the paper is organised as follows. Section II describes the video sequences used for the analysis as well as the extracted textural features. In Section III, the hypothesis of three video texture classes is verified. The proposed unsupervised classification models are trained and tested in Section IV. Finally, the conclusions are drawn in Section V.

II. FEATURE EXTRACTION AND SELECTION

II-A. Video Sequences used for Analysis

We employ a subset of a video data set that contains textures of different types from the HomTex database [5], [3]. HomTex is an annotated data set comprising homogeneous video textures of 256×256 spatial resolution at 25 and 60 fps. Using HomTex enables features to be extracted that better represent homogeneous video content. We selected a subset of HomTex clips that includes 15 static, 21 dynamic continuous and 25 dynamic discrete textures[1]. The selection criterion was "noise-free" features that best characterize the texture classes. Since some of the HomTex sequences exhibit acquisition noise (e.g. related to camera focus or compression noise), we visually inspected all the sequences and aimed at obtaining a "clean" subset.

Labelling of the sequences was based on the following definitions:

Dynamic Continuous: spatially irregular texture, with no clear structure, moving as a continuum e.g. water, deformable surfaces or smoke.

Dynamic Discrete: spatially regular or irregular texture that consists of perspectively moving independent discernible parts or structures, e.g. straws or leaves moving in a blowing wind.

Static: rigid texture that exhibits perspective motion, typically a moving solid object or a static background shot with camera motion, e.g. camera panning over a carpet.

II-B. Textural Features

There exists a rich literature describing different textural features designed for image and video analysis purposes. In this paper, we specifically extract features that cover the basic characteristics of video texture that relate to encoding difficulty, i.e. spatial diversity, coarseness and motion. All features were computed based solely on uncompressed video sequences. A total of 49 features with their statistics were extracted as explained below and summarized in Table I.

The *Gray Level Co-occurrence Matrix* (GLCM) [12] is a commonly used spatial textural feature that expresses the intensity contrast of neighbouring pixels in a frame, thus capturing the degree of coarseness and directionality of the texture. We computed the mean values of the GLCM descriptors, namely the mean contrast (abbreviated as con), the mean correlation (cor), the mean homogeneity (hom), the mean energy (enr), and the mean entropy (ent) on a frame level. Then, we calculated the mean value and standard deviation (std) of all these descriptors over the whole sequence.

The *Normalized Cross-Correlation* (NCC) [13], which is commonly used in image processing applications for spatial similarity purposes, is used as in [4] as a spatio-

[1]http://vilab.blogs.ilrt.org/files/2017/02/clean_HomTex_sequences.zip

Table I: List of features and notations.

Feature	Keywords
GLCM	$\text{meanGLCM}_{\text{con}}$, $\text{stdGLCM}_{\text{con}}$, $\text{meanGLCM}_{\text{cor}}$, $\text{stdGLCM}_{\text{cor}}$, $\text{meanGLCM}_{\text{hom}}$, $\text{stdGLCM}_{\text{hom}}$, $\text{meanGLCM}_{\text{enr}}$, $\text{stdGLCM}_{\text{enr}}$, $\text{meanGLCM}_{\text{ent}}$, $\text{stdGLCM}_{\text{ent}}$
NCC	NCC_{mean}, NCC_{std}, NCC_{skw}, NCC_{kur}, NCC_{ent}
ALPD	$\text{ALPD}_{\text{mean}}$, ALPD_{std}
NLP	NLP_{mean}, NLP_{std}, NLP_{skw}, NLP_{kur}
TC	$\text{meanTC}_{\text{mean}}$, $\text{stdTC}_{\text{mean}}$, $\text{meanTC}_{\text{std}}$, $\text{stdTC}_{\text{std}}$, $\text{meanTC}_{\text{skw}}$, $\text{stdTC}_{\text{skw}}$, $\text{meanTC}_{\text{kur}}$, $\text{stdTC}_{\text{kur}}$, $\text{meanTC}_{\text{entr}}$, $\text{stdTC}_{\text{entr}}$
OF	$\text{meanOF}_{\text{mag}}$, $\text{stdOF}_{\text{mag}}$, $\text{meanOF}_{\text{or}}$, stdOF_{or}, $\text{meanOF}_{\text{curl}}$, $\text{stdOF}_{\text{curl}}$, $\text{meanOF}_{\text{ang}}$, $\text{stdOF}_{\text{ang}}$, $\text{meanOF}_{\chi^2\text{mag}}$, $\text{stdOF}_{\chi^2\text{mag}}$, $\text{meanOF}_{\chi^2\text{or}}$, $\text{stdOF}_{\chi^2\text{or}}$, $\text{meanOF}_{\text{covVx}}$, $\text{stdOF}_{\text{covVx}}$, $\text{meanOF}_{\text{covVy}}$, $\text{stdOF}_{\text{covVy}}$, $\text{meanOF}_{\text{covVxVy}}$, $\text{stdOF}_{\text{covVxVy}}$

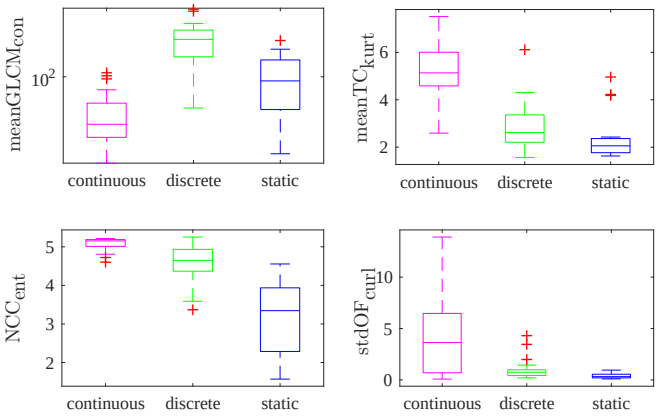

Fig. 1: Example boxplots of four extracted features per class selected during the visual assessment.

temporal feature by capturing the peaks of cross-correlation between successive frames. By using an overlapping 32×32 template between successive frames we computed the mean, the standard deviation, the skewness (skw), the kurtosis (kur) and the entropy of the peaks. All these statistics were averaged over the sequence.

As a measure of the coarseness, we used the *Average Local Peak Distance* (ALPD) in the third level of the discrete wavelet transform [14], [15]. For all sequences, we computed the average local peak distance on a frame level and then took the mean and standard deviation over all frames.

If we assume that each frame is a distorted version of its previous neighbour, we can use *Normalized Laplacian Pyramids* (NLPs) [16] to express this level of "distortion". We computed the mean, the standard deviation, the skewness, the kurtosis and the entropy of the NLP between successive frames and then we averaged these statistics over all frames.

In order to express how easy or difficult one frame can be predicted from its previous temporal neighbour, we used the *Temporal Coherence* (TC), as in [4]. We computed the mean, the standard deviation, the skewness, the kurtosis and the entropy on a frame level between successive frames. We took the mean and standard deviation of all these statistics over all pairs of successive frames.

The final feature employed is related to *Optical Flow* (OF), which has been computed based on Farneback's method [17]. OF descriptors and statistics are very important for the characterization of dynamic textures, since dynamic continuous textures exhibit different patterns of OF compared to dynamic discrete textures. We extracted the OF vectors together with the following statistics: mean and standard deviation of magnitude (mag), mean and standard deviation of orientation (or), mean and standard deviation of curl, mean and standard deviation of angular velocity (ang), χ^2-distance metric of magnitude and orientation histograms, mean and standard deviation of covariance of horizontal OF vectors (covVx), mean and standard deviation of covariance of vertical OF vectors (covVy), mean and standard deviation of covariance of horizontal and vertical OF vectors (covVxVy).

II-C. Feature Selection

The features extracted together provide a detailed description of the characteristics of the different texture classes. However, in order to build a robust classifier, reducing the dimensionality of the feature space and the selection of a suitable subset are important factors [18]. Therefore, we have used a feature selection method based on *Random Forest* (RF) models [19]. RFs are a popular type of machine learning technique due to the fact that they are robust even with high dimensional data and also capture both linear and non-linear relationships. They also employ feature ranking techniques, which can be used for feature selection. In order to rank the features, the model computes the mean decrease in the Gini impurity index every time it decides to split a node in order to grow each decision tree. Subsequently, those features which produce the highest mean decrease in the Gini impurity index are highly likely to result in better classification.

Fig. 2 illustrates the resulting ranking of the considered features. It is noticeable that the first 13 features are ranked in an almost linearly decreasing order while the rest of the features exhibit a uniform distribution with significant mean decrease Gini impurity values. According to the mean decrease Gini index, this means that the latter features (36 out of 49) are not likely to increase the classification accuracy.

It should be noted that the mean decrease Gini index ranking is not the only possible solution to feature selection. Other methods could produce different feature rankings with a similar impact on classification. Hence, for feature selection, we did not solely rely on the RF ranking, but also on a careful examination of the features per class. An example of the visual assessment of the feature selectivity is illustrated in Fig. 1. It is obvious that the depicted features have different ranges of the second and third quartiles per class, however they overlap in the minimum and maximum values.

Particularly, from the total of 49 features, we selected the 13 highly ranked from Fig. 2 and two more from the visual feature examination process. This particular subset of features achieved the highest accuracy compared to other combinations, while having the smallest cardinality. This subset of features is listed below: 1. $\text{meanGLCM}_{\text{con}}$, 2. stdOF_{or}, 3. $\text{meanGLCM}_{\text{cor}}$, 4. NCC_{ent}, 5. $\text{meanTC}_{\text{kur}}$, 6. $\text{meanTC}_{\text{skw}}$, 7. $\text{meanGLCM}_{\text{hom}}$, 8. $\text{meanTC}_{\text{entr}}$, 9. $\text{meanTC}_{\text{std}}$, 10. NCC_{mean}, 11. $\text{meanGLCM}_{\text{entr}}$, 12. NCC_{std}, 13. $\text{stdGLCM}_{\text{hom}}$, 14. $\text{stdOF}_{\text{cav}}$, and 15. $\text{stdOF}_{\text{curl}}$. These are adequate to both cluster and classify the video sequences in the following Sections.

III. VALIDATION OF NUMBER OF TEXTURE CLASSES THROUGH CLUSTERING

In this section, first we justify that it is meaningful to categorise content into three classes based on the extracted features. We show this graphically and then employ an unsupervised learning algorithm that estimates the optimal number of clusters (classes).

In order to visualize the data set in the feature space, we represent the data as an undirected graph as illustrated in Fig. 3. In this graph, each video sequence is a node with the edges connecting the N nearest neighbours of each sequence using the squared Euclidean distance across the selected feature set. For Fig. 3, we have used $N = 7$ and the colours magenta, green and cyan to distinguish static, dynamic discrete and dynamic continuous texture sequences, respectively (according to their annotations). This graph identifies three distinct clusters in the data set. As

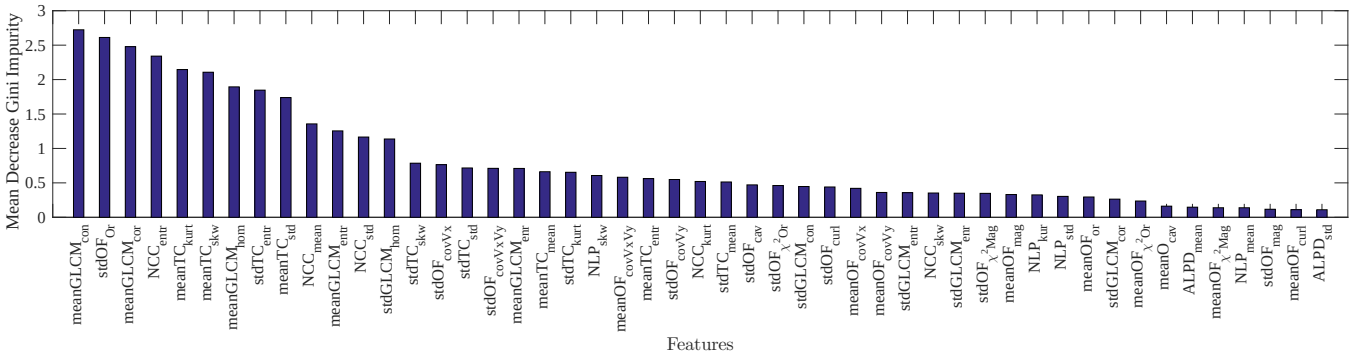

Fig. 2: Feature ranking based on RF model that uses the Mean Decrease Gini Impurity metric.

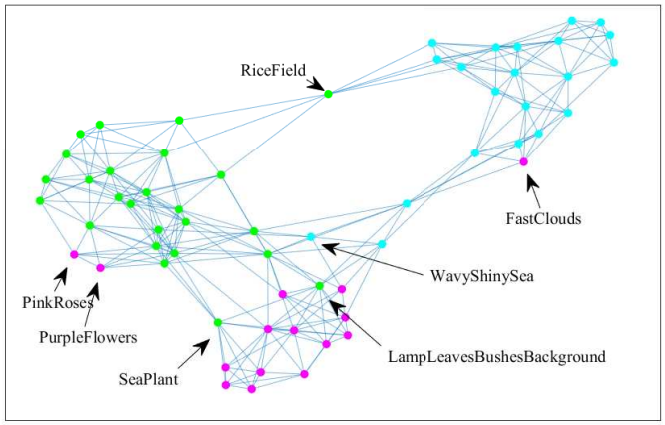

Fig. 3: Undirected graph representation of the sequences: static textures (magenta), dynamic discrete (green) and dynamic continuous (cyan).

annotated on the graph, a small number of sequences are "mislocated", namely FastClouds, SeaPlant, PinkRoses and PurpleFlowers. The sequence FastClouds is labelled as static and is located in the class of dynamic continuous. A possible explanation for this is that, despite its slow motion, it should have been annotated as continuous due to its continuously changing shape. For the SeaPlant sequence its slow motion results in locating it to the static sequences. Regarding the two static sequences, PinkRoses and PurpleFlowers, that are located within the dynamic discrete cluster, the reason is that their strong spatial characteristics in combination with the camera panning results in a distance close to other correctly clustered sequences. Also, there are some sequences that appear adjacent to a different class, such as RiceField and WavyShinySea. The reason for this is that these sequences exhibit some characteristics representative of another class. For example, WavyShinySea is a sequence with slowly moving sea water with reflections that look like small discrete structures and, therefore, it is located between the dynamic continuous and static classes. For the RiceField sequence,

Table II: Confusion Matrix for EM clustering on the extracted features.

		Clusters		
		Continuous	Discrete	Static
Labels	Continuous	19	0	2
	Discrete	1	21	3
	Static	1	2	12

the explanation is that, although the nature of its content is dynamic discrete, the camera captures its motion as dynamic continuous due to the camera distance. Thus, its feature distance places it between the two dynamic classes. The dynamic discrete LampLeavesBushesBackground is located in the static class because a significant part of each frame in this sequence is static with only a smaller part moving perspectively with high velocity.

Following visual inspection of the texture classes, we employ EM clustering. EM iteratively identifies the maximum likelihood estimates of the model parameters. The advantage of EM is that it can identify the optimal number of clusters for the input features [20]. Applying EM on our data set, the optimal number of clusters is three. In Table II, we show the confusion matrix of the produced EM clusters. This confirms that most of the clustered sequences are grouped in accordance to the data set annotations. Some of the sequences that are not aligned with their labelled class are the "mislocated" sequences from Fig. 3. For the clustering of the textures into classes, we have also evaluated the k-Means algorithm with $k = 3$ and the *Density-based Spatial Clustering of applications with Noise* (DBSCAN) [18]. Both of these algorithms produce almost the same results as EM. Both algorithms form clusters with high cohesion. This is confirmed by the clustering validity indices, i.e. average Silhouette, Purity, *Normalized Mutual Information* (NMI) and *Adjusted Rand Index* (ARI) [18], as reported in Table III.

Table III: Performance metrics of the clustering algorithms.

Algorithm	Silhouette	Purity	NMI	ARI
EM	0.5797	0.8524	0.5637	0.6179
k-Means	0.4544	0.8361	0.5210	0.5658
DBSCAN	0.4223	0.8196	0.501	0.5307

IV. CLASSIFYING VIDEO TEXTURE

Following validation of the hypothesis that video texture can be classified into three classes based on their spatio-temporal features, this Section proposes a classification scheme based on the same features that were used for clustering. First, a classifier is built using the labels for the three classes resulting from expert annotations of HomTex (Section IV-A). The second classifier ignores the expert annotations and uses labels produced by clustering the HomTex sequences based on their HEVC HM encoding statistics and validating the correlation of the spatio-temporal features of the uncompressed content to the compression performance (Section IV-B).

For both tasks, we used the clean annotated subset of 61 video sequences from HomTex. The 15 selected features were computed at a sequence level and, to avoid overfitting, a five-fold cross-validation was used. Several supervised learning methods with different configurations (e.g. different kernels, different distance functions, etc.) were tested, including *Nearest Neighbours* (NN), RF and *Support Vector Machines* (SVM) with different kernels. The best performance was obtained for two classifiers, the SVM with a quadratic kernel and a NN using euclidean distance. Given the robustness of SVMs in high dimensional classification problems [18], we propose to use the SVM for future work in the area.

IV-A. Using the Experts' Annotations Labelling

The classification accuracy achieved in this classification task is on average 95.1% and the results from the five-fold cross-validation of the proposed SVM classifier are shown in Table IV. This shows the confusion matrix of the proposed classifier and the *Area Under the Curve* (AUC) values per class. As can be observed, the AUC values are very high (maximum AUC value equals to 1) for all three classes showing a high accuracy in the prediction per class. The confusion matrix shows that there are only three misclassified video sequences, which are FastClouds, WavyShinnySea and Moving Pattern. The reasons for the incorrect classification for FastClouds and WavyShinnySea are the same as explained in Section III, Fig. 3. For the static sequence MovingPattern, a possible explanation is that although its motion is slow, it moves perspectively.

IV-B. Using Labelling based on HM Encoding Statistics

It is worth highlighting the fact that the expert-generated texture annotations used in HomTex are imperfect in the aforementioned classification systems, since they are compounded by a semantic bias. In most cases, this is due to the existence of a mixture of either spatial characteristics or temporal characteristics in a given sequence. For example, in WavyShinnySea, despite the reflections on the surface having structural characteristics, this sequence is still considered dynamic continuous by experts. Thus, this particular sequence (it could be argued) is a mixture of both dynamic continuous and dynamic discrete texture. In cases like this, the experts have annotated the sequences influenced by the semantics of the video content. For this reason, we considered important to create labels only based on the encoding performance as expressed through the encoding statistics.

Table IV: Confusion Matrix and AUC values of the proposed SVM classifier that uses expert annotations.

		Predicted classes			
		Continuous	Discrete	Static	AUC
Labels	Continuous	20	0	1	0.99
	Discrete	0	25	0	0.99
	Static	1	1	13	0.98

The labels used for this classification were created by applying EM clustering on the HM statistics for the case of Random Access and for quantization level equal to 25 from [3]. The HM statistics (37 in total) included statistics from prediction modes, reference indices, partitioning, residual, bit allocation, distortion and motion vectors at CTU level for different frame types (I, B and P). The classification accuracy achieved in this classification task is on average 90.2% and the confusion matrix from the five-fold cross-validation of the proposed SVM classifier is reported in Table V. As can be observed, the AUC values are lower compared to the respective values of the previous classifier for all three classes. The confusion matrix shows that there are six misclassified video sequences, which are: WavyShinySea, LampLeavesBushesBackground, SeaPlant, Stairs, FlowingRiver, and FastClouds. The main reason for these misclassifications is that these sequences lie near the borders of the areas of the different clusters. This means that these sequences are likely to exhibit characteristics of both these classes.

Table V: Confusion Matrix and AUC values of the proposed SVM classifier that uses the HM categorisation as labels.

		Predicted classes			
		Continuous	Discrete	Static	AUC
Labels	Continuous	20	1	2	0.91
	Discrete	1	21	1	0.97
	Static	1	0	14	0.96

Both classifications described above contribute in the understanding of video texture and its compression performance. The second classification has validated that spatio-temporal features extracted from uncompressed video textures can be directly used to accurately predict the expected encoder performance. This is particularly important as these same spatio-temporal features could be used for building a recommendation system for fast and optimised encoder configuration.

V. CONCLUSION

This paper has presented a study of different texture types in the context of HEVC HM encoder performance. It has clearly demonstrated that this is correlated with the spatio-temporal features extracted from uncompressed content. It has been verified using EM clustering that it is valid to categorize textures into three main classes: dynamic continuous, dynamic discrete and static from both the perspective of the input uncompressed video and from the encoding statistics associated with the output compressed video. Also, after selecting a subset of 15 extracted spatio-temporal features, an SVM classifier with high accuracy on homogeneous texture content was proposed. This classifier has been proved to be able to predict the HM encoding behaviour based on the spatio-temporal features. Since textures represent challenging video content from the point of view of compression, our in-depth analysis of texture classes and their behaviour, can be exploited for the optimisation of both the current and future encoding technologies.

VI. REFERENCES

[1] M. A. Papadopoulos, D. Agrafiotis, and D. R. Bull, "On the performance of modern video coding standards with textured sequences," in *International Conference on Systems, Signals and Image Processing (IWSSIP)*, Sept 2015, pp. 137–140.

[2] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, Dec 2012.

[3] M. Afonso, A. Katsenou, F. Zhang, D. Agrafiotis, and D. R. Bull, "Video Texture Analysis based on HEVC Encoding Statistics," in *2016 Picture Coding Symposium (PCS)*, Dec 2016.

[4] A. Katsenou, M. Afonso, D. Agrafiotis, and D. R. Bull, "Predicting Video Rate-Distortion Curves using Textural Features," in *2016 Picture Coding Symposium (PCS)*, Dec 2016.

[5] M. Afonso, A. Katsenou, F. Zhang, D. Agrafiotis, and D. R. Bull, "Homogeneous Video Texture Dataset (HomTex)," 2016, https://data.bris.ac.uk/data/datasets/1h2kpxmxdhccf1gbi2pmvga6qp/.

[6] P. Saisan, G. Doretto, Y. N. Wu, and S. Soatto, "Dynamic texture recognition," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, vol. 2, pp. II–II.

[7] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *International Journal of Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003.

[8] P. Renaud, S. Fazekas, and M. J. Huiskes, "DynTex: a Comprehensive Database of Dynamic Textures," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1627–1632, 2010.

[9] K. Derpanis and R. Wildes, "Spacetime texture representation and recognition based on a spatiotemporal orientation analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1193–1205, 2012.

[10] X. Qi, C.-G. Li, G. Zhao, X. Hong, and M. Pietikinen, "Dynamic texture and scene classification by transferring deep image features," *Neurocomputing*, vol. 171, pp. 1230–1241, 2016.

[11] C. Theriault, N. Thome, and M. Cord, "Dynamic Scene Classification: Learning Motion Descriptors with Slow Features Analysis," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[12] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, Nov 1973.

[13] J. P. Lewis, "Fast template matching," in *Vision interface*, 1995, vol. 95, pp. 15–19.

[14] P. Salembier and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, John Wiley and Sons, Inc., New York, NY, USA, 2002.

[15] M. M. Subedar and L. J. Karam, "A no reference texture granularity index and application to visual media compression," in *2015 IEEE Intern. Conf. on Image Processing (ICIP)*, Sept 2015, pp. 760–764.

[16] V. Laparra, J. Ball, A. Berardino, and E.P. Simoncelli, "Perceptual image quality assessment using a normalized laplacian pyramid," in *Electronic Imaging 2016*. SPIE, 2016, pp. 1–6.

[17] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian Conference on Image Analysis*. Springer, 2003, pp. 363–370.

[18] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Third Edition*, Academic Press, Inc., Orlando, FL, USA, 2006.

[19] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Advances in Neural Information Processing Systems*, 2013, pp. 431–439.

[20] U. D. Gupta, V. Menon, and U. Babbar, "Detecting the number of clusters during expectation-maximization clustering using information criterion," in *Second IEEE International Conference on Machine Learning and*

Computing (ICMLC), 2010, pp. 169–173.